We claim:

1. A method for automatically extracting information from a semi-structured subsequent document having a document type, the document type comprising design and structural characteristics of the document, the subsequent document containing terms having respective data elements, the method comprising:

   a) providing at least one initial document of the document type that also contains terms having respective data element values;

   b) designing an extraction template for the terms of the document type of each initial document;

   c) tagging the terms of each initial document according to the extraction template;

   d) providing selection criteria for each tagged term of each initial document;

   e) building a text mining term model based on each tagged term and its respective analyzed adjacent content;

   f) optimizing the text mining term model using a regression optimization algorithm;

   g) providing the subsequent semi-structured document; and

   h) using the optimized text mining term model to automatically extract information from the subsequent document.

2. The method of claim 1, in which the text mining term model is optimized repeatedly.

3. The method of claim 1, in which the text mining term model is optimized periodically.

4. The method of claim 1, in which the text mining term model is optimized by analyzing newly tagged and subsequent documents to determine whether the regression optimization algorithm may be improved.

5. The method of claim 1, further comprising retaining specific information about a set of similar documents having a common feature, and using a pattern represented by the specific information to search for text in the subsequent document.

6. The method of claim 5, in which the set of similar documents is a set of initial documents.

7. The method of claim 5, in which the set of similar documents is a set of subsequent documents.

8. The method of claim 5, in which the common feature comprises at least one of the structure of the similar documents, the pages of the similar documents, tables of the similar documents, table layouts of similar documents.

9. The method of claim 5, in which the common feature is the identity of the company to which the set of similar documents pertains.

10. The method of claim 1, in which the tagging of the terms of each initial document according to the extraction template is performed manually.

11. The method of claim 1, in which the tagging of the terms of each initial document according to the extraction template is performed manually and comprises creation of an experience set and training of the text mining term model.

12. The method of claim 1, further comprising re-optimizing the text mining term model based on accuracy of the text mining term model when extracting information from any subsequent semi-structured document.

13. The method of claim 1, in which designing the extraction template for the terms of the document type of each initial document comprises optimization of invariants.

14. The method of claim 1, further comprising adding each initial document to a document repository prior to tagging each initial document to the extraction template.

15. The method of claim 1, in which the tagging of the terms of each initial document according to the extraction template is performed automatically by a pattern recognition process.

16. The method of claim 1, in which the text mining term model is optimized without human interaction.

17. The method of claim 1, further comprising adding each subsequent document to a document repository prior to using the text mining term model to automatically extract information from the subsequent document.

18. The method of claim 1, in which the text mining term model is built and used without previously defined rules regarding characteristics of data within the document.

19. The method of claim 1, in which the text mining term model is built and used without fixed criteria specifying prior knowledge of data within any document.

20. The method of claim 1, further comprising a quality control process that tests the text mining term model using a control group of subsequent documents having the same document type that have not been provided previously.

21. The method of claim 1, further comprising a quality control process that tests extracted data quality.

22. The method of claim 1, in which there is one text mining term model for each term.

23. The method of claim 1, in which each of the terms required for extraction is identified using a taxonomy of term names for each of the terms.

24. The method of claim 1, in which each of the terms required for extraction is identified using various attributes for each of the terms.

25. The method of claim 1, in which the method is implemented as a set of application programming interfaces invoked by a programming environment.

26. The method of claim 25, in which the programming environment is one of the group consisting of Java, C, C++, and Visual Basic.

27. The method of claim 25, in which the programming environment provides at least one of the initial document and subsequent document.

28. The method of claim 25, in which the programming environment uses the optimized text mining term model.

29. The method of claim 25, in which the programming environment receives information extracted from the subsequent document.

30. The method of claim 1, further comprising collection of metadata from at least one document.

31. The method of claim 30, in which the metadata comprises at least one of row and column header strings, footnote information, name of the document, date and time stamp data, and note or comment information.

32. The method of claim 30, further comprising linking the metadata to a source of the document from which it was collected.

33. The method of claim 1, further comprising using a wizard to produce the text mining term model.

34. The method of claim 1, further using a wizard to automatically create a decision tree to provide hierarchical selection criteria for determining the location of the information.

35. The method of claim 1, further comprising a wizard to schedule optimization of the text mining term model.

36. The method of claim 1, further comprising a wizard to simulate user interface actions required by input to the wizard.

37. The method of claim 1, further comprising converting data from documents of disparate formats into a uniform data format.

38. The method of claim 37, in which the data in the uniform data format is used while building the text mining term model.

39. A system for automatically extracting information from a semi-structured subsequent document having a document type, the document type comprising design and structural characteristics of the document, the subsequent document containing terms having respective data elements, the system comprising:

a) a source of at least one initial document of the document type that also contains terms having respective data element values;

b) an extraction template for the terms of the document type of each initial document;

c) means for tagging the terms of each initial document according to the extraction template;

d) means for providing selection criteria for each tagged term of each initial document;

e) a text mining term model based on each tagged term and its respective analyzed adjacent content;

f) an regression optimization algorithm that optimizes the text mining term model;

g) a source of the subsequent semi-structured document;

h) means for using the optimized text mining term model to automatically extract information from the subsequent document.

40. The system of claim 39, in which the text mining term model is optimized repeatedly.

41. The system of claim 39, in which the text mining term model is optimized periodically.

42. The system of claim 39, in which the text mining term model is optimized by analyzing newly tagged and subsequent documents to determine whether the regression optimization algorithm may be improved.

43. The system of claim 39, further comprising means for retaining specific information about a set of similar documents having a common feature, and using a pattern represented by the specific information to search for text in the subsequent document.

44. The system of claim 43, in which the set of similar documents is a set of initial documents.

45. The system of claim 43, in which the set of similar documents is a set of subsequent documents.

46. The system of claim 43, in which the common feature comprises at least one of the structure of the similar documents, the pages of the similar documents, tables of the similar documents, table layouts of similar documents.

47. The system of claim 43, in which the common feature is the identity of the company to which the set of similar documents pertains.

48. The system of claim 39, in which the tagging of the terms of each initial document according to the extraction template is performed manually.

49. The system of claim 39, in which the tagging of the terms of each initial document according to the extraction template is performed manually and comprises creation of an experience set and the training of the text mining term model.

50. The system of claim 39, further comprising means for re-optimizing the text mining term model based on accuracy of the text mining term model when extracting information from any subsequent semi-structured document.

51. The system of claim 39, in which the extraction template for the terms of the document type of each initial document comprises optimized invariants.

52. The system of claim 39, further comprising means for adding each initial document to a document repository prior to tagging each initial document to the extraction template.

53. The system of claim 39, in which the means for tagging of the terms of each initial document according to the extraction template automatically performs a pattern recognition process.

54. The system of claim 39, in which the text mining term model is optimized without human interaction.

55. The system of claim 39, further comprising means for adding each subsequent document to a document repository prior to using the text mining term model to automatically extract information from the subsequent document.

56. The system of claim 39, in which the text mining term model is built and used without previously defined rules regarding characteristics of data within the document.

57. The system of claim 39, in which the text mining term model is built and used without fixed criteria specifying prior knowledge of data within any document.

58. The system of claim 39, further comprising means for quality control of the text mining term model based on a control group of subsequent documents having the same document type that have not been provided previously.

59. The system of claim 39, further comprising means for quality control of extracted data quality.

60. The system of claim 39, in which there is one text mining term model for each term.

61. The system of claim 39, in which each of the terms required for extraction is identified using a taxonomy of term names for each of the terms.

62. The system of claim 39, in which each of the terms required for extraction is identified using various attributes for each of the terms.

63. The system of claim 39, in which the system comprises a set of application programming interfaces invoked by a programming environment.

64. The system of claim 63, in which the programming environment is one of the group consisting of Java, C, C++, and Visual Basic.

65. The system of claim 63, in which the programming environment provides at least one of the initial document and subsequent document.

66. The system of claim 63, in which the programming environment uses the optimized text mining term model.

67. The system of claim 63, in which the programming environment receives information extracted from the subsequent document.

68. The system of claim 39, further comprising means for collecting metadata from at least one document.

69. The system of claim 68, in which the metadata comprises at least one of row and column header strings, footnote information, name of the document, date and time stamp data, and note or comment information.

70. The system of claim 68, further comprising means for linking the metadata to a source of the document from which it was collected.

71. The system of claim 68, further comprising a wizard that produces the text mining term model.

72. The system of claim 68, further comprising a wizard that automatically creates a decision tree to provide hierarchical selection criteria for determining the location of the information.

73. The system of claim 68, further comprising a wizard that schedules optimization of the text mining term model.

74. The system of claim 39, further comprising a wizard that simulates user interface actions required by input to the wizard.

75. The system of claim 39, further comprising means for converting data from documents of disparate formats into a uniform data format.

76. The system of claim 75, in which the data in the uniform data format is used while building the text mining term model.

77. A workflow for extracting information from a semi-structured subsequent document having a document type, the document type comprising design and structural characteristics of the document, the subsequent document containing terms having respective data elements, the workflow comprising:

a) receiving at least one initial document of the document type that also contains terms having respective data element values;

b) designing an extraction template for the terms of the document type of each initial document;

c) tagging the terms of each initial document according to the extraction template;

d) providing selection criteria for each tagged term of each initial document;

e) building a text mining term model based on each tagged term and its respective analyzed adjacent content;

f) optimizing the text mining term model using a regression optimization algorithm;

g) receiving the subsequent semi-structured document;

h) using the optimized text mining term model to extract information from the subsequent document.

78. The workflow of claim 77, further comprising repeatedly optimizing the text mining term model.

79. The workflow of claim 77, further comprising using a document repository for any initial document.

80. The workflow of claim 79, in which the document repository supports at least one of the processes selected from the group consisting of checking

out documents, manually tagging term values, and correcting errors from the auto-extraction process.

81. The workflow of claim 79, in which the document repository displays results of a quality control process.

82. The workflow of claim 77, in which the text mining term model is optimized repeatedly.

83. The workflow of claim 77, in which the text mining term model is optimized periodically.

84. The workflow of claim 77, further comprising analyzing newly tagged and subsequent documents to determine whether the regression optimization algorithm may be improved.

85. The workflow of claim 77, further comprising retaining specific information about a set of similar documents having a common feature, and using a pattern represented by the specific information to search for text in the subsequent document.

86. The workflow of claim 85, in which the set of similar documents is a set of initial documents.

87. The workflow of claim 85, in which the set of similar documents is a set of subsequent documents.

88. The workflow of claim 85, in which the common feature comprises at least one of the structure of the similar documents, the pages of the similar documents, tables of the similar documents, table layouts of similar documents.

89. The workflow of claim 85, in which the common feature is the identity of the company to which the set of similar documents pertains.

90. The workflow of claim 77, in which the tagging of the terms of each initial document according to the extraction template is performed manually and comprises creation of an experience set and the training of the text mining term model.

91. The workflow of claim 77, further comprising re-optimizing the text mining term model based on accuracy of the text mining term model when extracting information from any subsequent semi-structured document.

92. The workflow of claim 77, in which the tagging of the terms of each initial document according to the extraction template is performed manually.

93. The workflow of claim 77, in which designing the extraction template for the terms of the document type of each initial document comprises optimization of invariants.

94. The workflow of claim 77, further comprising adding each initial document to a document repository prior to tagging each initial document to the extraction template.

95. The workflow of claim 77, in which the tagging of the terms of each initial document according to the extraction template is performed automatically by a pattern recognition process.

96. The workflow of claim 77, in which the text mining term model is optimized without human interaction.

97. The workflow of claim 77, further comprising adding each subsequent document to a document repository prior to using the text mining term model to automatically extract information from the subsequent document.

98. The workflow of claim 77, in which the text mining term model is built and used without previously defined rules regarding characteristics of data within the document.

99. The workflow of claim 77, in which the text mining term model is built and used without fixed criteria specifying prior knowledge of data within any document.

100. The workflow of claim 77, further comprising a quality control process that tests the text mining term model using a control group of subsequent documents having the same document type that have not been provided previously.

101. The workflow of claim 77, further comprising a quality control process that tests extracted data quality.

102. The workflow of claim 77, in which there is one text mining term model for each term.

103. The workflow of claim 77, in which each of the terms required for extraction is identified using a taxonomy of term names for each of the terms.

104. The workflow of claim 77, in which each of the terms required for extraction is identified using various attributes for each of the terms.

105. The workflow of claim 77, in which the method is implemented as a set of application programming interfaces invoked by a programming environment.

106. The method of claim 105, in which the programming environment is one of the group consisting of Java, C, C++, and Visual Basic.

107. The method of claim 105, in which the programming environment provides at least one of the initial document and subsequent document.

108. The method of claim 105, in which the programming environment uses the optimized text mining term model.

109. The method of claim 105, in which the programming environment receives information extracted from the subsequent document.

110. The workflow of claim 77, further comprising collection of metadata from at least one document.

111. The workflow of claim 110, in which the metadata comprises at least one of row and column header strings, footnote information, name of the document, date and time stamp data, and note or comment information.

112. The workflow of claim 110, further comprising the metadata to a source of the document from which it was collected.

113. The workflow of claim 110, further comprising using a wizard to produce the text mining term model.

114. The workflow of claim 77, further using a wizard to automatically create a decision tree to provide hierarchical selection criteria for determining the location of the information.

115. The workflow of claim 77, further comprising a wizard to schedule optimization of the text mining term model.

116. The workflow of claim 77, further comprising a wizard that simulates user interface actions required input to the wizard.

117. The workflow of claim 77, further comprising converting data from documents of disparate formats into a uniform data format.

118. The workflow of claim 117, in which the data in the uniform data format is used while building the text mining term model.